Fairness and Biases in AI Algorithms and Interfaces

Seul Lee[a]

[a]University of California, Los Angeles, Department of Information Studies, United States

seul@g.ucla.edu

ABSTRACT

This workshop will explore various biases embedded in AI algorithms and interfaces across online platforms, including digital archives and libraries, social media, search engines, and AI-powered services like ChatGPT. Participants will critically engage with issues surrounding information credibility, transparency in content selection and appraisal, the accuracy and integrity in representation, and the intricate dynamics of information authority, format, and editorial oversight. The workshop aims to equip participants with practical strategies to identify, assess, and address biases inherent in various media, empowering them to navigate and evaluate online information across platforms with informed judgment.

Below is a detailed timeline outlining the activities and delivery methods for the workshop:

Section 1. Biases in social media (appx. 60 minutes)

1-1. Lecture on biases in information on social media (appx. 20 minutes)
   1-1.1. Introduction (appx. 10 minutes)
   1-1.2. Interactive lecture and discussion on fake news, mis/disinformation (appx. 10 minutes)
1-2. Hands-on workshop (appx. 20 minutes)

Break (10 minutes)

Section 2. Biases in digital archives, libraries, and search engines (appx. 50 minutes)

2-1. Lecture and case studies on biases in digital archives, libraries, and search engines (appx. 40 minutes)
2-2. Small group discussion (appx. 10 minutes)
Break (10 minutes)

Section 3. Biases in ChatGPT (appx. 50 minutes)

3-1. Lecture on biases in ChatGPT (appx. 30 minutes)

3-2. Large group discussion and reflection (appx. 20 minutes)

The workshop will begin with a 10-minute introduction, welcoming participants and outlining the goals of the session. The workshop will kick off with an interactive introductory lecture focused on various potential biases in AI algorithms. The first part, which will last about 10 to 20 minutes, will address recognizing biases in social media. Participants will analyze these biases by identifying fake news in provided articles and reflecting on their own everyday information practices on social media. After this informative segment, participants will engage in hands-on workshops aimed at applying their knowledge to recognize the types of biases they may encounter during their online information-seeking activities. They will participate in hands-on activities lasting 20 to 30 minutes, using different online tools to help identify such biases. Following the interactive hands-on workshops, participants will have the opportunity to share their insights and findings regarding the biases they discovered.

The second section, which will explore biases in digital archives, digital libraries, and search engines through case studies, will last approximately one hour. It will include a 40-minute lecture that covers different types of biases in digital archives, libraries, and search engines, illustrated through case studies. After the lecture, participants will reflect on the biases presented for 10 minutes in a group discussion, sharing their thoughts and insights.

Following participants' reflections on these biases, the last part of the session will focus on potential biases in ChatGPT. It will begin with a 30-minute lecture on how ChatGPT generates responses, emphasizing potential biases that could arise at each layer or step of the process. By examining a step-by-step analysis of how ChatGPT generates its answers, participants will deepen their understanding of the biases and errors inherent in the information generated by ChatGPT and its complex operational dynamics. Following this lecture, participants will engage in discussion for 20 minutes, working in small groups to identify biases and discussing necessary educational changes for themselves and their students. The workshop will conclude with a final reflection, where participants will share insights gained throughout the session and discuss how to apply their learning moving forward. The workshop will be wrapped up in a brief 10-minute summary of key points, encouraging participants to implement the strategies discussed in their own information evaluation practices.

This workshop is designed for beginner-level students and researchers, as well as those interested in incorporating critical digital literacy principles into their teaching. It is open to people with diverse backgrounds and levels of expertise. No prior knowledge is required, allowing participants to embark on their learning journey with a fresh perspective and receptive attitude. The goal of this workshop is to enhance critical thinking skills related to the credibility of information while unpacking the complexities of information curation, representation, and editorial control in today's intricate digital landscape. Participants will gain a more nuanced understanding of how information is selected, represented, and potentially biased across various platforms, including social media, digital archives and libraries, search engines, and AI-powered services like ChatGPT. Through this workshop, participants will learn to effectively navigate the complexities of their information practices and develop the ability to critically assess the credibility of the online information they consume.

Through engaging lectures and hands-on activities, participants will learn to identify and evaluate the inherent biases present in various media channels. They will explore practical strategies for assessing the credibility of information, recognizing potential mis/disinformation, and critically analyzing the content they encounter in their everyday information-seeking activities. By the end of the workshop, participants will be equipped with the necessary tools to navigate the digital information landscape safely and effectively, empowering them to make informed decisions not only for themselves but also in guiding others, such as children, students, or peers, in developing their own critical digital media literacy skills. This comprehensive approach aims to enhance participants' ability to engage thoughtfully with information, ultimately fostering a more informed and critically aware community.

ALISE RESEARCH TAXONOMY TOPICS

Information Literacy; Information Ethics; Archival Arrangement and Description; Artificial Intelligence; Pedagogy.

AUTHOR KEYWORDS

Information Biases; Algorithmic Biases; Information Credibility; Digital Archives; Search Engines; Generative AI; ChatGPT.

DOI: https://doi.org/10.21900/j.alise.2024.1701

References

Barocas, S., Hardt, M. & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities.* The MIT Press. https://fairmlbook.org/

Drucker, J., (2013). Reading Interface. *PMLA, 128*(1), 213-220. http://www.jstor.org/stable/23489280

Gilliland-Swetland, L. J. (1991) The Provenance of a Profession: The Permanence of the Public Archives and Historical Manuscripts Traditions in American Archival History. *The American Archivist, 54*(2), 160–175. http://www.jstor.org/stable/40293549

Goldman, E. (2006). Search Engine Bias and the Demise of Search Engine Utopianism. *8 Yale J. L. & Tech*. 188. https://digitalcommons.law.scu.edu/cgi/viewcontent.cgi?article=1112&context=facpubs

Hardt, M. (2014, September 26). *How Big Data is Unfair.* Medium. https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de

Yakel, E. (2003). Archival representation. *Archival Science, 3*, 1-25. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/41831/10502_2004_Article_5139967.pdf?sequence=1